Solutions for review problems

Modules 7-12

Stat 251

updated 2021

(1) Ithaca, NY is located in upstate New York and averages around 37" of rain each year, with standard deviation of 3.25". Rainfall in Ithaca, NY tends to follow an approximately normal distribution. Say that a student attends Conrnell University in Ithaca and is there for 4 years.

(a) Define the Central Limit Theorem.
The sampling distribution of the sample mean will be approximately normal with mean μ and standard deviation (also known as standard error) $\frac{\sigma}{\sqrt{n}}$, provided the sample size, n, is large ($n \geq 30$)

(b) Describe the sampling distribution of the sample mean of rainfall in Ithaca, NY. Include the mean of the sampling distribution of the sample mean and the standard deviation of the sampling distribution of the sample mean (standard error).
The distribution should be approximately normal with mean 37 and standard error $= \frac{3.25}{\sqrt{4}} = 1.625$ so $\overline{X} \sim N(37, 1.625)$

(c) What is the probability that during the 4 years, we see an average less than 32"?
Now use $z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$ $P(\bar{x} < 32) = P(Z < \frac{32-37}{\frac{3.25}{\sqrt{4}}}) = P(Z < -3.08) = 0$

(d) What is the 93^{rd} percentile for average precipitation? $z_{top7\%} = z_{.93} = 1.48$ now solve for $\overline{X}$ where $\overline{X} = z(\sigma/\sqrt{n}) + \mu \Rightarrow \overline{X} = (1.48)(1.625) + 37 = 39.4$

(2) Ithaca, NY is located in upstate New York and averages around 37" of rain each year, with standard deviation of 3.25". Rainfall in Ithaca, NY tends to follow an approximately normal distribution. Say that a student attends Conrnell University in Ithaca and is there for 7 years (undergraduate and graduate degrees).

(a) Describe the sampling distribution of the total rainfall in Ithaca, NY. Include the total of the sampling distribution and the standard deviation of the sampling distribution of the total (standard error)
$\hat{\tau} \sim N(n\mu, \sqrt{n}\sigma) \Rightarrow \hat{\tau} \sim N(259, 8.5986918)$

(b) What is the probability that during the 7 years, we see a total precipitation of less than 220"?
$P(\hat{\tau} < 220) = P(Z < \frac{220-259}{8.6}) = P(Z < -4.5348837) = 2.8127114 \times 10^{-6} \approx 0$

(c) What is the 93^{rd} percentile for total precipitation?
$z_{93\%} = 1.48 \Rightarrow 1.48 = \frac{\hat{\tau}-259}{8.6} \Rightarrow \hat{\tau} = 271.7$"

(3) A survey of purchasing agents from 250 randomly selected industrial companies found that 25% of the buyers reported higher levels of new orders in January than in earlier months.

(a) Describe the sampling distribution of the proportion of buyers in the US with higher levels of new orders in January. Include the mean of the sampling distribution and the standard deviation of the sampling distribution (standard error)
$\hat{p} \sim N(p, \sqrt{pq/n}) \Rightarrow \hat{p} \sim N(0.25, 0.0273861)$

(b) What is the probability that the sample proportion is more than 26%?

$P(\hat{p} > 0.26) = P(Z > \frac{0.26-0.25}{0.0274}) = P(Z > 0.36) = 1 - P(Z < 0.36) = 1 - 0.6406 = 0.3594$

(c) What is the probability that the sample proportion is less than 20%?
$P(\hat{p} < 0.2) = P(Z < \frac{0.2-0.25}{0.0274}) = P(Z < -1.82) = 0.0344$

(4) A company that manufactures coffee for use in commercial machines monitors the caffeine content in its coffee. The company randomly selected 50 samples of coffee every hour from its production line and determines the caffeine content. From historical data, the caffeine content is known to have a standard deviation of 7.1 mg. During one 1-hour period, a random sample of 50 had a sample mean of 110 mg.

(a) The caffeine content should usually be 107 mg. Is there sufficient evidence that the mean caffeine content is more than the usual amount? Hypotheses: $H_0 : \mu = 107$ and $H_a : \mu > 107$ Assumptions: random (yes), independence (yes because random), normality (met because $n \geq 30$)
$t = 4.375$, $df = 49$, $pvalue =$`3.17e-05`$= 3.17 \times 10^{-5} \approx 0 \leq \alpha(0.05) H_0$ is rejected. There is sufficient evidence the true average caffeine content is more than the usual amount of 107 mg.

(b) What kind of error could have been made? Define the error and explain it in the context of the scenario.
Since we rejected H_0, the only kind of error we could have made was a Type I error, α. It is defined as $P(Reject\ H_0 | H_0\ true)$, rejecting the null hypothesis when the null hypothesis is true. We think the caffeine content is more than 107 mg but it is not.

(c) Estimate μ, the true average caffeine content of the coffee, with 95% confidence. Interpret.
CI: $109.5083, 113.7704$ mg. We are 95% confident the true mean caffeine content is between 109.5 and 113.8 mg.

(d) *Potential extra credit question*: Suppose that the company would prefer a margin of error (bound) for the next batch to be 1.15. What sample size would be needed to get a bound of 1.15, while maintaining 95% confidence using $z^\star$ for the critical value?
$n = \left(\frac{z^\star \sigma}{bound}\right)^2 = \left(\frac{(1.96)(7.1)}{1.15}\right)^2 = 146.4310442$ and since it is a sample size, we always round up (regardless of normal rounding conventions). So $n \approx 147$

(5) It is thought that more than 70% of all faults in transmission lines are caused by lightning. In a random sample of 200 faults from a large data base, 151 are due to lightning.

(a) Is there sufficient evidence that the proportion of faults in transmission due to lightning strikes is different from 70%? Conduct a hypothesis test. Hypotheses: $H_0 : p = 0.7$ $H_a : p \neq 0.7$ Assumptions: random (yes), independence (yes because random), normality (met because $n \geq 60 = 200$)
$t = 1.804$, $df = 199$, $pvalue = 0.07275 \nleq \alpha(0.05) \therefore H_0$ cannot be rejected.
Conclusion: Since we did not reject H_0, the proportion of faults in transmission due to lightning strikes is still right around 70% (it is not different from 70%).

(b) What kind of error could have been made? Define the error and explain it in the context of the scenario. Since we did not reject H_0, the only kind of error we could have made was a Type II error (β). It is defined as $P(Fail\ to\ Reject\ H_0 | H_0\ false)$, not rejecting a false hypothesis. We think the faults in transmission due to lightning is not different from the usual 70%, when it would be different from the 70%.

(c) Estimate p, the true proportion of faults in transmission due to lightning strikes, with 95% confidence. Interpret.
CI: $(0.6948788, 0.8151212) = (69.49\%, 81.51\%)$. We are 95% confident the true proportion of faults of transmission due to lightning strikes is between 69.49% and 81.51%.

(d) *Potential extra credit question*: Suppose that next sample should have a bound of 3%. What sample size would be needed to get a bound of 3%, while maintaining 95% confidence using $z^\star$ for the critical value?
$n = \hat{p}\hat{q}\left(\frac{z^\star}{bound}\right)^2 = (0.755)(0.245)\left(\frac{1.96}{0.03}\right)^2 = 789.5555111 \Rightarrow n \approx 790$

(6) In 1882 Michelson measured the speed of light (usually denoted as c in Einstein's equation $E = mc^2$). He reported the results of 23 random trials with a mean of 29756.22 km/sec and standard deviation of 107.12 km/sec. [Note that the actual speed of light is actually 299,792,458 metres per second but remember, these results are from experiments done in *1882*; use the units of measurement that Michelson used (km/sec).]

(a) Suppose previous experiments of Michelson found that the speed of light was 29750 km/sec. Is there sufficient evidence from his experiments that the speed of light is significantly different from the previous result of 29750? Let $\alpha = 0.02$. Hypotheses: $H_0 : \mu = 750$ and $H_a : \mu \neq 750$
Assumptions: random (yes), independence (yes because random), normal (assume)
$t = 1.3576$, $df = 22$, $pvalue = 0.1884 \nleq \alpha(0.02) \therefore$ $H_0$is not rejected.
Conclusion: Since H_0 was not rejected, there is not sufficient evidence the true average speed of light is different than previous experiments' results of 29750 km/sec.

(b) What kind of error could have been made? Define the error and explain it in the context of the scenario.
Since we did not reject H_0, the only kind of error we could have made was a Type II error (β). It is defined as $P(Fail\ to\ Reject\ H_0 | H_0\ false)$, not rejecting a false hypothesis. We think the speed of light is 29750 but it is different than previous results.

(c) Estimate μ, the true speed of light with 98% confidence. Interpret.
CI: (29722.12, 29843.66) km/sec. We are 98% confident the true speed of light is between 29722.12 and 29843.66 km/sec.

(7) *pvalues*: Find the *pvalue* of the following scenarios and state whether or not the null hypothesis should be rejected for each one. We can reject H_0 if $pvalue \leq \alpha$.

(a) $H_0 : \mu = 5$ vs. $H_a : \mu < 5$, $z_{calc} = -1.55$. $pvalue = P(Z < z_{calc}) = P(Z < -1.55) = 0.0606$. $0.0606 \nleq \alpha(0.05)$ so H_0 is not rejected

(b) $H_0 : \mu = 5$ vs. $H_a : \mu > 5$, $z_{calc} = 1.55$, $\alpha = 0.10$
$pvalue = P(Z > z_{calc}) = 1 - P(Z < z_{calc}) = 1 - P(Z < 1.55) = 1 - 0.9394 = 0.0606$. $0.0606 \leq \alpha(0.10)$ so H_0 is rejected

(c) $H_0 : \mu = 5$ vs. $H_a : \mu \neq 5$, $z_{calc} = -1.55$
Since $z_{calc} < 0$, $pvalue = 2P(Z < z_{calc}) = 2P(Z < -1.55) = 2(0.0606) = 0.1212$. $0.1212 \nleq \alpha(0.05)$ so H_0 is not rejected

(8) *Difference of 2 independent means test and CI* $\mu_1 - \mu_2$: Researchers speculate that drivers who do not wear a seatbelt are more likely to speed than drivers who do wear one. A random sample of 40 drivers was taken. In the experiment, the people were clocked to see how fast they were driving (mph) and then were stopped to see whether or not they were wearing a seatbelt.

(a) Is there sufficient evidence that the average speed for non-seatbelt wearers differs from those drivers that do wear a seatbelt? Let $\alpha = 0.10$ $H_0 : \mu_1 = \mu_2$ vs. $H_a : \mu_1 \neq \mu_2$
$t = 2.771$, $df = 37.524$, $pvalue = 0.01565 \leq \alpha(0.10) \therefore$ H_0 is rejected. There is a significant difference in the speeds of drivers who do not wear seatbelts as compared to those who wear seatbelts.

(b) Estimate the true difference in means of the speeds of drivers who do not wear seatbelts as compared to those who wear seatbelts with 90% confidence and interpret.
CI: (2.053407, 10.244202). We are 90% confident the true difference in means of the speeds of drivers who do not wear seatbelts as compared to those who wear seatbelts is between 2.05 and 10.24 mph (those that do not wear seatbelts travel between 2.05 and 10.24 mph faster than those that wear a seatbelt)

(c) State the kind of error that could have been made. **DESCRIBE IT IN CONTEXT OF THE DATA**. Since H_0 was rejected, a Type I error (α) could have been made, thinking that there is a difference in the speed of drivers when there isn't.

(9) *Paired t-test and CI (dependent samples)* μ_D: Many freeways have service (or logo) signs that give information on attractions, camping, lodging, food, and gas services prior to off-ramps. An article reported that in one investigation, six sites along Virginia interstate highways where service signs are posted were selected randomly. For each site, crash data was obtained for a three-year period before

distance information was added to the service signs for a one-year period afterward.

(a) Is there sufficient evidence that there is a decrease in accidents after the signs added distance information?

In this case, since differences are calculated as $after - before$ and we want to know if there is a decrease *after*, the hypothesized difference should be < 0. It could have been done the other way around and all it would do is make $\overline{X}_d$ positive, s_d would not change, the t score would be positive, the rejection region would change to the upper tail, and the result would not change.

$H_0 : \mu_D = 0$ vs. $H_a : \mu_D < 0$
$t = -0.15141$, $df = 5$, $pvalue = 0.4428 \nleq \alpha(0.01) \therefore$ H_0 cannot be rejected.
There is no significant difference in accidents before and after the signage, or the signage did not seem to be effective.
(b) Estimate the true mean difference in accidents before and after the signage change with 99% confidence and interpret.
CI: $(-88.69426, 82.27443)$. Because the interval contains 0, the only conclusion we can make is that there is no significant difference in accidents before or after the sign change. We are 99% confident the true mean difference of accidents before and after the signage change is between -89 and 82 But logically this does not make sense. All this tells us is that there is no difference in accidents (just like the hypothesis test in part a because if the hypothesized value is in the CI ($\mu_D = 0$), then you cannot reject H_0). (c) State the kind of error that could have been made. **DESCRIBE IT IN CONTEXT OF THE DATA**. With H_0 not rejected, a Type II error (β) could have been made, thinking that there is no change in accidents but there could be a decrease in the accidents.

(10) *Difference of two proportions test and CI*: A hospital administrator suspects that the delinquency rate in the payment of hospital bills has increased over the past year. Hospital records show that the bills of 48 of 1284 persons admitted in the month of April have been delinquent for more than 90 days. This number compares to 34 of 1002 persons admitted during the same month one year ago. (a) Is there sufficient evidence that there is an increase in delinquency rate in the payment of hospital bills over the last year? $H_0 : \pi_1 = \pi_2$, $H_a : \pi_1 > \pi 2$
test statistic $t = 0.4426$, $df = 2194.2$, $pvalue = 0.671 \nleq \alpha(0.05) \therefore$ H_0 is not rejected. There is no evidence that the current years' delinquency rate has increased from last year.
(b) Estimate the true difference of proportions of delinquent bills over the last year with 95% confidence and interpret CI: $(-0.01183959, 0.01874168) \approx (-1.18\%, 1.87\%)$. With 95% confidence, the true difference in proportions of delinquent rate from this year to last year is between -1.18% and 1.87%. With 0 being in the CI, there is no significant difference in the years' rates
(c) State the kind of error that could have been made. **DESCRIBE IT IN CONTEXT OF THE DATA** Since H_0 was not rejected, a Type II error could have been made. We think the delinquency rates have not increased but they have.

(11) *Book Mediums*: (not the psychic kind of medium) A professor of an introductory college class uses an open-source textbook for the class. Of interest is the proportions of students that will either purchase a hard copy, print the book online, or just use the downloaded PDF format to read on a device. From earlier semesters, 60% bought a hard copy of the book, 25% printed it online, and 15% used a downloaded PDF format on their devices. At the end of the semester, the professor asks the students to complete a survey and indicate what format of the book they used. Of the 126 students, 71 bought a hard copy, 30 printed it, and 25 downloaded PDF to use. (a) Is there evidence that the students used similar mediums for the book? (b) State the kind of error that could have been made. *Describe in context*

```
       type counts probs
1 Hard copy     71  0.60
2   Printed     30  0.25
3       PDF     25  0.15
```

Null Hypothesis
H_0 : Students use of books is 60% hard copy, 25% printed, 15% PDF (or $H_0 : p_1 = 0.6,\ p_2 = 0.25,\ p_3 = 0.15$)

Alternative hypothesis
H_a : H_0 is not true (students use of books are not the estimated percents as listed above)

Expected values

$$E_i = np_i$$

Here we can check to see all $E_{ij} \geq 5$

Results
Test Statistic: $\chi^2 = 2.3201$, $df = 2$, $pvalue = 0.3135 \nleq \alpha(0.05) \therefore$ H_0 is not rejected.

Conclusion (in context)
We did not reject H_0, indicating that the students are using the different book mediums as expected (similar to other semesters).

Error
We did not reject H_0 so a type II error could have been made. We think that students are using the different mediums of books as expected but they are not.

(11) *Independence Test* χ^2: In Star Trek fandom, there is a running joke that characters on the show who wear a red shirt are doomed, just another statistic. Shirt colors can be only blue, gold, or red; fatalities can be only dead or alive. (a) Is there sufficient evidence determine whether there is an association

between shirt color and deaths? (b) State the kind of error that could have been made. **DESCRIBE IT IN CONTEXT OF THE DATA**

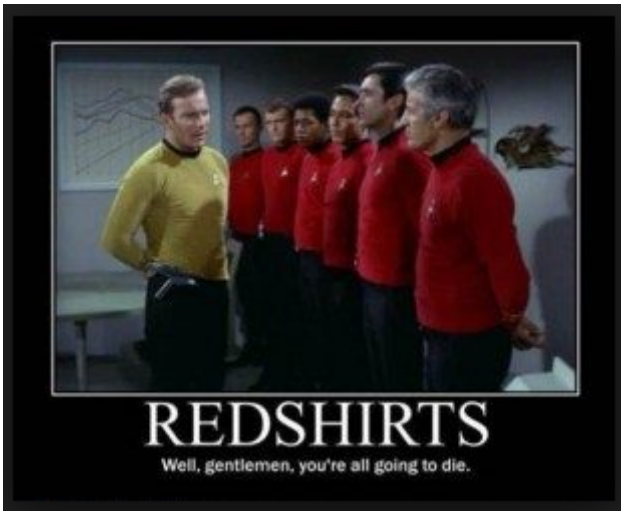


Figure 1: Red Shirt of Doooom

```
        Shirt.Colour
Survival Blue Red Gold Total
   Alive  129 215   46   390
   Dead     7  24    9    40
   Total  136 239   55   430

Star Trek survival by shirt colour
```

Null Hypothesis
H_0 : Shirt colour and survival on Star Trek are independent
H_a : H_0 is not true (Shirt colour and survival on Star Trek are dependent)

Expected values

The expected counts are given and all $E_{ij} \geq 5$

Results
Test statistic $\chi^2 = 6.1886$, $pvalue = 0.04531 \leq \alpha(0.05) \therefore$ H_0 is rejected.

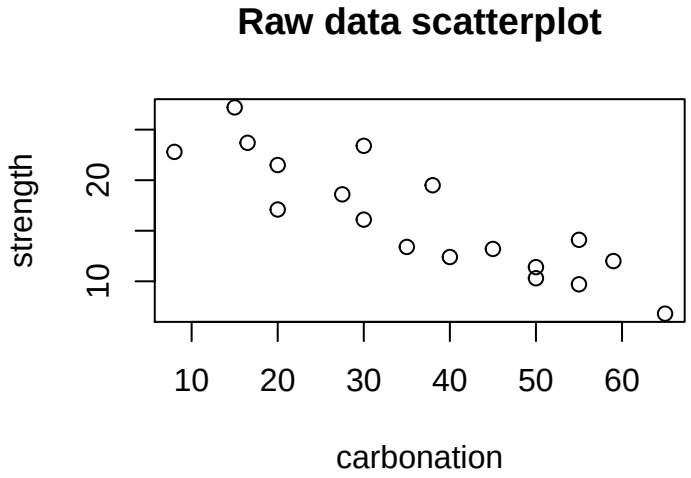
Conclusion (in context)
We rejected H_0 so that tells us that the dreaded red shirt does mean you are less likely to survive an episode of Star Trek (survival is dependent on shirt colour).

Error
We rejected H_0 so a type I error could have been made. We think that survival depends on shirt colour when shirt colour makes no difference in survival.

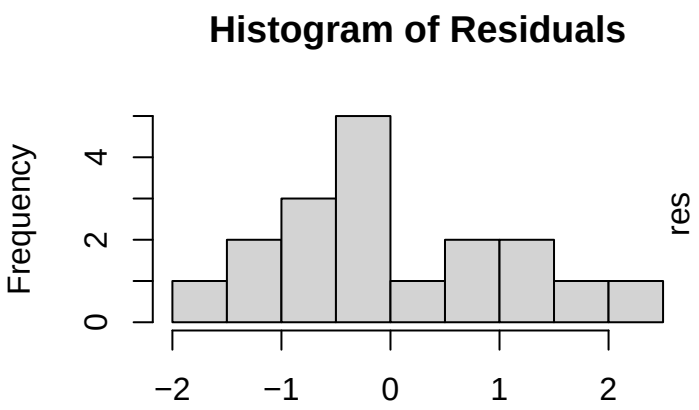
(12) *SLR (simple linear regression)*

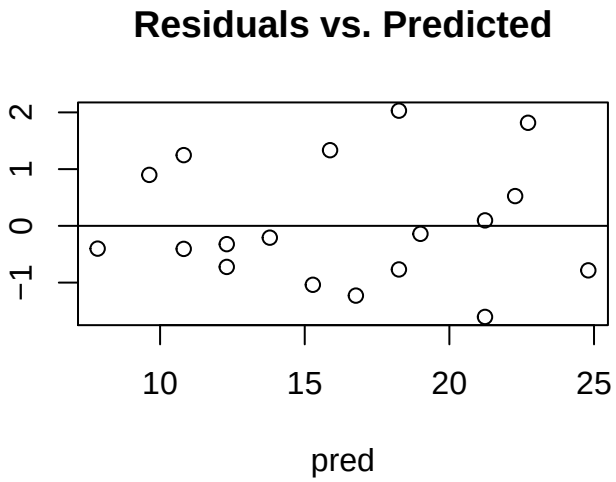
Corrosion of steel reinforcing bars is the most important durability problem for reinforcing structures. Carbonation of concrete results form a chemical reaction that lowers the pH value by enough to initiate corrosion of the rebar. Data on the carbonation depth (mm) and strength (MPa) for a sample of core specimens was taken from a particular building, and all the regression output is provided. We are interested in modeling the strength. (a) State the (population) model equation and define its components (b) Looking at the raw data scatterplot, does it appear as if there is a linear relationship? Positive or negative slope? (c)

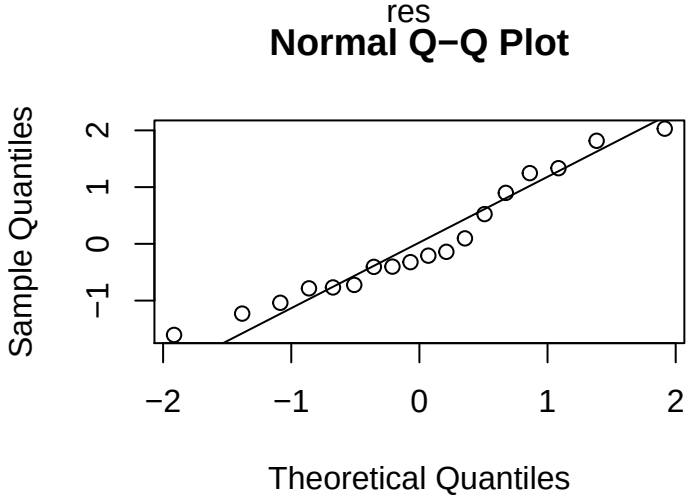
State the regression equation using the provided output. Using the regression equation, estimate the strength when the carbonation depth is 8 mm and estimate it again when the depth is 20 mm. (d) Calculate the residuals for both of your estimates in part c. The observed value for 8 mm is 22.8 MPa ($(8, 22.8)$) and for 20 mm is 17.1 MPa ($(20, 17.1)$). (e) Interpret slope and intercept *in context* of the data. If something does not make sense in context, state it and describe why. (f) What is the coefficient of determination, R^2? List the value and interpret in context. (g) What is the correlation, r? List the value annd interpret in context. (h) Is the slope significant? Conduct hypothesis test; include hypotheses, test statistic, *df*, *pvalue*, results, and conclusion. (i) List assumptions of regression (words or symbols) and check assumptions (assumptions 1, 2, and 4). Briefly describe how they are/are not met. (j) Using parts b, f, g, h, and i, is this a good model? Reference those to verify your claim (briefly describe).




```
Call:
lm(formula = strength ~ carbonation)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1317 -2.0043 -0.7488  2.1366  5.1439

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.18294    1.65135  16.461 1.88e-11 ***
carbonation -0.29756    0.04116  -7.229 2.01e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.864 on 16 degrees of freedom
Multiple R-squared:  0.7656,    Adjusted R-squared:  0.7509
F-statistic: 52.25 on 1 and 16 DF,  p-value: 2.013e-06
```







Population model

$$y = \beta_0 + \beta_1 x + \epsilon_i$$

y: response variable
β_0: intercept when $x = 0$
β_1: slope (change in y due to 1 unit increase in x)
x: explanatory variable
ϵ_i: residual (random error) term

There appears to be a negative linear relationship between carbonation and strength

regression equation
The regression equation is

$$\hat{y} = 27.18 - 0.298x$$

$\hat{y}|_{x=8} = 27.18 - 0.298(8) = 24.796$
$\hat{y}|_{x=20} = 27.18 - 0.298(20) = 21.22$

residuals
$((8, 22.8))$ and for 20 mm is 17.1 MPa $((20, 17.1))$

$$e_i = y - \hat{y}$$

$e_{x=8} = 22.8 - 24.8 = -2 < 0 \therefore$ the estimate was an overestimate
$e_{x=20} = 17.1 - 21.2 = -4.1 < 0 \therefore$ the estimate was an overestimate

Interpretation of slope and intercept:
Slope: a one mm increase in the carbonate depth will reduce (because the slope is negative) the strength by 0.296 MPa.
Intercept: when depth is 0 mm, the strength is 27.183 MPa. Even though $x = 0$ is not in our dataset, this could make logical sense in context, if the strength here is the base strength.

Coefficient of Determination R^2:
Listed on the regression output as `Multiple R-squared`: 0.7656; this means that 76.56% of the variation (in the response) can be explained by the model. Since $76.56\% \geq 60\%$, this is good and the model is good at explaining most of the variation.

Correlation r:
Not listed in the regression output. $r = \pm\sqrt{R^2}$ with the sign depending on the sign of the slope. Since the slope here is neagtive, then $r = -\sqrt{R^2} = -\sqrt{0.7656} = 0.8749857$. Since this is "close" to -1, then we have a moderately strong, negative, *linear* relationship between carbonation (x) and strength (x). In fact, as carbonation increases, strength decreases (lower carbonation=higher strength and vice versa).

Null Hypothesis
$H_0 : \beta_1 = 0$

Alternative hypothesis
$H_a : \beta_1 \neq 0$

Test Statistic, df, pvalue
$t = -7.229$, $df = 16$, and $pvalue = 2.01\text{e-}06 = 2.01 \times 10^{-6} \approx 0$

Reject H_0 if $pvalue \leq \alpha(0.05)$. Since $0 \leq 0.05$, H_0 is rejected, meaning the slope is significant.

Error
We rejected H_0 so a type I error could have been made. We think that the slope is significant when it is not, meaning if we have made this error, any estimations done will basically mean nothing.

There is a strong, negative linear relationship between carbonation and strength, R^2 is good (more than 60%), r is decent (-0.87), and the slope is significant so this should be a good model.

Assumptions:
(1) $E(\epsilon_i) = 0$: the mean of the residuals is ≈ 0. The histogram of residuals is centered around 0 so the assumptions is **met** (2) $V(\epsilon_i) = \sigma_\epsilon^2$: the variance of the residuals is constant (same for all values of y). There is random scatter on the `Residuals vs. Predicted` plot, with no meaningful pattern so the assumption is **met** (3) $Cov(\epsilon_i, \epsilon_j) = 0$: independence of residuals. No check for this; **assume met** (4) $\epsilon_i \sim N(0, \sigma_\epsilon^2)$: residuals have an approximate normal distribution with mean 0 and constant variance. The histogram of residuals is roughly normal (normal enough); the qqplot shows that most points are on the $y = x$ line. This assumption is **met**